DETECCIÓN AUTOMÁTICA DE MARCADORES DISCURSIVOS DEL ESPAÑOL

UNA APLICACIÓN CON XFST

KOZA, Walter Adrián
Grupo Infosur – Universidad Nacional de Rosario
kozawalter@opendeusto.es

Fecha de recepción:
14 de febrero de 2012
Fecha de revisión:
13 de mayo de 2012
Fecha de aceptación:
25 de junio de 2012

Resumen: El reconocimiento automático de los marcadores discursivos es una tarea fundamental en la lingüística informática. No obstante, uno de los problemas que se suscita tiene que ver con los casos de ambigüedad, es decir, construcciones idénticas que actúan como marcadores discursivos en algunos contextos y en otros no. Para casos como estos, la puntuación cumple un rol esencial. Aquí, se propone un método de detección automática, tomando como base las funciones gramaticales de la coma y las hipótesis de desambiguación presentadas por Prada (2001). Para ello, se recurrió a Xfst, un autómata de estados finitos desarrollado por *Xerox*, que, entre sus posibilidades, incluye el comando «*tokenize*» con el que se reconocen y balizan los marcadores discursivos. El artículo se organiza de la siguiente manera: en primer lugar, se presenta una descripción de los marcadores discursivos y la clasificación que de ellos presentan Martín Zorraquino y Portolés (1999). Luego se traen a colación los antecedentes sobre el tratamiento de estas construcciones en la lingüística computacional. En tercer lugar, se describe la implantación en máquina y los resultados obtenidos. Finalmente, en cuarto lugar, se presentan las conclusiones derivadas de la investigación.

Palabras clave: Marcadores discursivos – análisis automático – Xfst – balizado.

Résumé: La reconnaissance automatique des marqueurs discursifs est une tâche fondamentale dans la linguistique informatique. Toutefois, l'un des problèmes qui se pose est liée à des cas d'ambiguïté, c'est-à-dire constructions identiques qui agissent comme marqueurs discursifs dans certains contextes mais ne le font pas dans d'autres. Aux cas comme ceux ci, la

---

Los marcadores del discurso en español y en chino mandarín, Francisco J. RODRÍGUEZ MUÑOZ (ed.), Almería: Universidad de Almería, 2012 (= *Philologica Urcitana. Revista Semestral de Iniciación a la Investigación en Filología,* vol. 7, Septiembre 2012, ISSN: 1989-6778), 59-74

ponctuation joue un rôle essentiel. Ici, on propose une méthode de détection automatique, sur la base des fonctions grammaticales de la virgule et les hypothèses de desambiguïté présentées par Prada (2001). Pour cela, on a utilisé Xfst, un programme automatique d'états finis développé par *Xerox*, lequel, entre ses possibilités, comprend la commande «tokenize» avec laquelle on reconnait et on balise les marqueurs discursifs. L'article s'organise de la manière suivante: tout d'abord, on présente une description des marqueurs discursifs et une de ses classifications présentée par Martín Zorraquino et Portolés (1999). Ensuite on apporte les antécédents sur le traitement de ces constructions dans la linguistique computationnelle. En troisième lieu, on décrit l'implantation en machine et les résultats obtenus. Finalement, en quatrième lieu, on expose les conclusions résultantes de la recherche.

Mots-clés: Marqueurs discursifs – analyse automatique – Xfst – baliser.

1. INTRODUCCIÓN

En este artículo, presento la descripción y una propuesta de reconocimiento automático de los marcadores discursivos. Para ello, adopto como base los lineamientos presentados en mi tesis doctoral sobre el análisis de las funciones gramaticales de la coma[1] dentro del marco de la lingüística computacional (Koza, 2011).

Así pues, mostraré una aplicación de la herramienta informática Xfst (Xerox Finite State Tools) (Beesley y Kartunnen, 2003), para el análisis y balizamiento de marcadores discursivos del español. Este *software* fue desarrollado por *Xerox* y usado por *Xerox Research Centre Europe* (XRCE, Grenoble, Francia) y *Palo Alto Research Center* (PARC, California, USA) y otros centros mundiales de investigación lingüística. Xfst permite trabajar los datos lingüísticos en forma independiente de la máquina algorítmica y no requiere que el investigador posea conocimientos de informática.

Guillot (2005: 65) describe su modo de análisis de la siguiente manera:

> La aplicación se presenta como una implementación de autómatas de estados finitos para producir análisis morfológicos y generación. Incluye *lexc*, un lenguaje declarativo de alto nivel usado para especificar lexicones de lengua natural. La sintaxis de este lenguaje ha sido diseñada para facilitar la definición de la estructura morfotáctica, tratamiento de grandes irregularidades y miles de formas base de una lengua natural. Los archivos fuentes declarativos pueden introducirse con un editor de textos planos como el *notepad* de *Windows* o el emacs de *Linux*.

Las herramientas que utiliza este programa son las siguientes:

- Tokenizadores de estado finito: ejecutan la segmentación del texto de acuerdo con la información morfosintáctica almacenada.
- Analizadores morfológico/generadores de estado finito (también denominados transductores lexicales).
- Desambiguadores de estado finito o taggers: examinan los tokens ambiguos en su contexto sintáctico y les asignan una única etiqueta o *tag* como «nombre» o «verbo».
- Analizadores sintácticos de estado finito o *parser*: reconocen un sintagma, lo etiquetan y lo ponen entre corchetes.

Para el trabajo de reconocimiento automático de los marcadores discursivos, es preciso analizar la relación que se establece entre ellos y la puntuación. Pues, los signos de

[1] Dicha tesis fue dirigida por la Dra. Zulema Solana y mis estudios de doctorado se financiaron con una beca de CONICET.

puntuación constituyen un mecanismo de organización de la lectura y permiten delimitar las llamadas unidades textuales (Nunberg, 1990):

- *Párrafo*: dado por punto y aparte.
- *Enunciado textual*: dado por punto y seguido.
- *Cláusula textual*: dada por punto y coma.
- *Enunciado oracional*: dado por dos puntos.
- *Sintagma*: dado por la coma.

La segmentación textual es una fase necesaria para una gran variedad de tareas realizadas en el tratamiento automático; ya sea para el análisis sintáctico, el resumen automático, el filtrado de textos, etcétera. En mi tesis doctoral, establecí propuestas de segmentación a partir de la detección automática de las funciones de la coma. Opté por analizar este signo dado que es el más complejo y el que mayor número de funciones gramaticales presenta. En primer lugar, cotejé las clasificaciones que, sobre este signo, proponen Alcoba (2000), Simone (1991), Figueras (2001) y el *Diccionario Panhispánico de Dudas* (RAE, 2005). Luego, establecí una clasificación propia, basada en criterios gramaticales y, a la vez, acorde con el trabajo computacional. A continuación, dispongo dicha clasificación.

Función indicadora

Mediante esta función, se indican los elementos de las enumeraciones simples (1) y las elipsis (2).

(1) Compró pan**,** vegetales y carne.

(2) Juan lee el periódico; María**,** una revista.

Función delimitadora

En este caso, la coma no indica, sino que delimita construcciones incidentales de diversos tipos (3, 4 y 5) y alteraciones en el orden regular (6).

(3) Juan**,** el marido de Luisa**,** acaba de perder el empleo.

(4) El profesor**,** que siempre estaba atento**,** reprendió a los alumnos que copiaban.

(5) No se preocupe**,** estimado amigo**,** yo me voy a ocupar del asunto personalmente.

(6) En el día de ayer**,** Juan y María fueron al cine.

Función desambiguadora

Mediante la coma se pueden desambiguar construcciones que pueden presentar más de un significado (7 y 8).

(7) No debes hacerlo.

(8) No**,** debes hacerlo.

En relación con el último ítem, la coma sirve, en algunos casos, para determinar si lo que encontramos en una construcción es o no un marcador discursivo. Por ejemplo, pueden observarse las diferentes funciones que cumplen las expresiones «por un lado» y «por otro lado» en (9) y (10):

(9) **Por un lado,** se prevé una «redistribución de los suministros de gas» que tienen contratados las centrales térmicas. (…) Y, **por otro lado**, a partir de 2007 se procederá a la desregulación total de los suministros y cualquier usuario residencial podrá contratar la provisión de energía con el generador o comercializador que más le convenga (*Clarín*, 29/04/04).

(10) La verdad va **por un lado**. La política va **por otro lado**.

En el primero de los ejemplos, se trata de marcadores discursivos estructuradores de la información cuya función es asociar la información de los dos párrafos, dada por la ilación «por un lado», «por otro lado». En cambio, en (10) las mismas construcciones actúan como complementos verbales.

En esta ocasión, se propone un método de detección automática de aquellos marcadores discursivos que requieren de coma para evitar ambigüedades. Esta investigación continúa los planteamientos desarrollados en Koza (2011), allí se describe el método propuesto con la herramienta informática Smorph (Aït-Mokthar, 1998).

El artículo se organiza de la siguiente manera:

En primer lugar, se presenta una descripción de los marcadores discursivos y la clasificación que de ellos presentan Martín Zorraquino y Portolés (1999). Luego se traerán a colación los antecedentes sobre el tratamiento de estas construcciones en la lingüística computacional, haciendo hincapié en el trabajo de Prada (2001). Posteriormente, se describe la implantación en máquina y los resultados obtenidos. Por último, se presentan las conclusiones derivadas de la investigación.

2. ACERCA DE LOS MARCADORES DISCURSIVOS

2.1. Consideraciones generales

Los marcadores discursivos aluden a un conjunto de términos que establecen relaciones entre segmentos textuales con el objetivo de guiar y ordenar los procesos de interpretación en la comprensión de textos. A comienzos de la década del setenta, dos disciplinas nacientes, la lingüística textual y la pragmática, se centraron en el estudio de estos términos en la medida en que confirmaban sus respectivos puntos de partida: la propuesta de romper las fronteras de la oración como límite último de los estudios del lenguaje (Portolés, 1998).

Martín Zorraquino y Portolés (1999: página) se refieren a ellos con la denominación de *marcadores del discurso* y los definen de la siguiente manera:

> Los *marcadores del discurso* son unidades lingüísticas invariables, no ejercen una función sintáctica en el marco de la predicación oracional –son, pues, elementos marginales– y poseen un cometido coincidente en el discurso: el de guiar, de acuerdo con sus distintas propiedades morfosintácticas, semánticas y pragmáticas, las inferencias que se realizan en la comunicación.

Una cuestión que conviene considerar es que los marcadores discursivos presentan una variada gama de particularidades y poseen una complejidad que escapa a todo intento exhaustivo de sistematización. No obstante, estos autores tratan de presentar una clasificación de ellos, basada en dos condiciones: a) los elementos agrupados deben compartir propiedades gramaticales homogéneas (los marcadores tratados se ajustan, en general, a las categorías tradicionales de los adverbios, las locuciones adverbiales y ciertas interjecciones); b) a su vez, sus características semánticas (la forma de significar o configurar su significado) deben ser las propias de los marcadores discursivos; esto es, que no presentan un contenido referencial o denotador, sino que muestran un significado de procesamiento.

Estas construcciones son de gran ayuda en el proceso de inferencias durante la comprensión lectora. Las inferencias son un conjunto de operaciones de razonamiento que realiza el lector en el momento de leer. Martín Zorraquino y Portolés solo van a considerar como marcadores a aquellas expresiones que no contribuyan al significado conceptual de los enunciados, sino que orienten y ordenen las inferencias que cabe obtener de ellos. Esto quiere decir que los marcadores contribuyen al procesamiento de lo que se comunica y no a la representación de la realidad comunicada.

2.2. Clasificación de los marcadores discursivos

Martín Zorraquino y Portolés distinguen cinco grupos de marcadores; el primero se denomina *estructuradores de la información* y se los utiliza para señalar la organización informativa de los discursos. Estos carecen de significado argumentativo y se dividen en:

- *Comentadores*: introducen un nuevo comentario («pues», «pues bien», «así las cosas»).
- *Ordenadores*: agrupan varios miembros del discurso como partes de un único comentario. Estos marcadores, por lo general, están basados en la numeración («primero», «segundo»), en lo espacial («por un lado»… «por otro lado», «por una parte»… «por otra parte») o en lo temporal («después», «luego», «finalmente»). Algunos de ellos forman pares correlativos que pueden estar seguidos por un tercer miembro también con ordenador: «por un lado/por otro (lado)», «por una parte/por otra (parte)», etcétera. Los ordenadores, a su vez se clasifican en tres tipos:
    1) *Marcadores de apertura*: abren una serie en el discurso («en primer lugar», «primeramente», «por un lado»).
    2) *Marcadores de continuidad*: indican que el miembro que acompañan forma parte de una serie de la cual no son el elemento inicial («en segundo/tercer/…/lugar», «por otra parte», «por otro lado»).
    3) *Marcadores de cierre*: señalan el fin de una serie discursiva («por último», «en último lugar», «finalmente»).
- *Digresores*: introducen un comentario lateral respecto de la planificación del discurso anterior («por cierto», «a todo esto», «a propósito»).

El segundo grupo de marcadores discursivos es el de los *conectores* que vinculan semántica y pragmáticamente a un miembro del discurso con otro anterior, de tal forma que el marcador guía las inferencias que se efectúan del conjunto de los dos miembros discursivos conectados. Pueden reconocerse tres grupos:

- *Conectores aditivos*: unen a un miembro anterior con otro de su misma orientación («además», «encima», «aparte», «incluso»).
- *Conectores consecutivos*: conectan a un consecuente con su antecedente («por tanto», «por consiguiente», «por ende», «en consecuencia»).

- *Conectores contraargumentativos*: eliminan algunas de las conclusiones que pudieran inferirse de un miembro anterior («en cambio», «por el contrario», «sin embargo», «no obstante»).

El tercer grupo de marcadores del discurso es el de los *reformuladores*. Estos presentan a un miembro del discurso como una expresión más adecuada de lo que se pretendió decir con un miembro precedente. Los reformuladores se subdividen en cuatro grupos, a saber:

- *Reformuladores explicativos*: presentan un nuevo miembro del discurso como una explicación anterior («o sea», «esto es», «es decir»).
- *Reformuladores rectificativos*: corrigen a un miembro discursivo anterior («mejor dicho», «mejor aún», «más bien»).
- *Reformuladores de distanciamiento*: privan de pertinencia al miembro discursivo anterior («en cualquier caso», «en todo caso», «de todos modos»).
- *Reformuladores recapitulativos*: introducen una recapitulación o conclusión de un miembro discursivo anterior o una serie de ellos («en suma», «en conclusión», «en definitiva», «en fin», «al fin y al cabo»).

El cuarto grupo es el de los *operadores argumentativos*. En este caso, el marcador condiciona, por su significado, las posibilidades argumentativas del miembro en que se incluye, sin relacionarlo con otro anterior. Pueden establecerse dos grupos de operadores argumentativos:

- *Operadores de refuerzo argumentativo*: su significado refuerza como argumento el miembro del discurso en el que se encuentra frente a otros posibles argumentos («en realidad», «en el fondo», «de hecho»).
- *Operadores de concreción*: muestran el miembro del discurso en el que se localizan como una concreción o un ejemplo de una generalización («por ejemplo», «en particular»).

Finalmente, el quinto grupo es el que remite a los marcadores *conversacionales*. Respecto de ellos, los autores aclaran que con esta división no se pretende determinar un límite estricto entre lo conversacional y lo no conversacional, puesto que:

> Todo discurso es, en esencia, dialógico y, de hecho, muchos de los marcadores que se han incluido en los grupos precedentes pueden aparecer también en la conversación; asimismo, bastantes marcadores conversacionales se emplean a menudo en los textos escritos (Martín Zorraquino y Portolés, 1999).

Sin embargo, la conversación constituye una situación comunicativa particular y con propiedades específicas que van a determinar o favorecer la presencia de ciertos marcadores. Los marcadores conversacionales se distribuyen en cuatro grupos:

- *De modalidad epistémica*: señalan el grado de certeza, de evidencia, etcétera, que el hablante atribuye al miembro o a los miembros del discurso con el que se vincula cada partícula («claro», «por lo visto», «desde luego»).
- *De modalidad deóntica*: indican diversas actitudes volitivas del hablante respecto del miembro o miembros del discurso en el que aquellos comparecen («bueno», «bien», «vale»).
- *Enfocadores de la alteridad*: orientan sobre la forma como el hablante se sitúa en relación con su interlocutor en la interacción comunicativa («hombre», «mira», «oye»).
- *Metadiscursivos conversacionales*: sirven para estructurar la conversación; es decir, para distinguir bloques informativos, por ejemplo, o para alternar o mantener los turnos de palabra, etcétera («bueno», «eh», «este»).

En el presente trabajo no tendrá cabida este grupo de marcadores, puesto que no contribuye al objetivo principal, que es el análisis de textos escritos.

3. SOBRE LOS MARCADORES DISCURSIVOS EN EL ÁMBITO DE LA LINGÜÍSTICA COMPUTACIONAL

Los marcadores discursivos son de gran utilidad en varias tareas del procesamiento del lenguaje natural, como el auto-resumen, la traducción automática, análisis sintáctico, etcétera, ya que aportan información muy rica sobre la estructura discursiva, con un bajo coste de procesamiento. No obstante, según señalan Alonso, Castellón y Padró (2002), la comunidad científica no ha llegado a un acuerdo respecto de su delimitación y caracterización:

> Esta falta de consenso se debe, por un lado, a la preeminencia de las aproximaciones de tipo deductivo, con un sesgo importante por una teoría subyacente y, por otro, a la subordinación de la mayor parte de caracterizaciones a una tarea computacional concreta, lo que suele conllevar soluciones *ad hoc*.

En relación con los antecedentes, entre los trabajos sobre los marcadores discursivos, pueden mencionarse los realizados por Knott y Dale (1995), quienes proponen mecanismos formales para la detección y sistematización de estas unidades. Posteriormente, Knott

(1996) aplica estos mecanismos para obtener y caracterizar un conjunto de unos 200 marcadores discursivos del inglés. A partir de la investigación anterior, Marcu (1997) desarrolla un sistema de análisis de la estructura retórica para el inglés basado en la información discursiva que obtiene de un conjunto de 400 marcadores discursivos.

Para Alonso, Castellón y Padró (2002), las falencias de los estudios mencionados radican en que no solucionan la creación de un listado extenso y no controvertido de los marcadores discursivos para uso computacional o cómo abordar la creación de estos recursos para otras lenguas.

Para el caso del español, estos autores presentan una construcción de un lexicón computacional de marcadores discursivos, implementado en un sistema de resumen genérico de extracción que puede funcionar autónomamente o en colaboración con otras técnicas o tipos de información. El sistema se compone de dos módulos que utilizan el lexicón:

- El *segmentador*, que detecta las unidades discursivas básicas.
- El *interpretador*, que pondera cada segmento según su relevancia discursiva.

Dicho lexicón fue implementado en un sistema de ayuda para el resumen automático. La primera etapa del trabajo fue realizada mediante métodos empíricos e implantación en un sistema de resumen. Posteriormente, presentan mejoras estructurales y de contenido mediante la aplicación de técnicas *clustering*.

Otra de las investigaciones realizadas sobre los marcadores discursivos es la realizada por Prada (2001), quien tomando como punto de partida la clasificación de Martín Zorraquino y Portolés, desarrolla una propuesta de implantación en máquina de una serie de reglas que permiten el reconocimiento automático de los marcadores. Los signos de puntuación cumplen un papel importante para solucionar problemas de ambigüedad como los descritos en (9) y (10), y Prada formula una serie de hipótesis de desambiguación haciendo hincapié en ellos. Tales hipótesis son las siguientes:

Hipótesis 1: Ordenadores

Se podrá considerar estructura ordenadora a aquella en la que cada marcador de la secuencia sea el primer elemento en una oración o en un párrafo y además debe estar seguida de una coma. En el caso de ocurrencias intraoracionales (fundamentalmente para los de continuidad), la coma o una conjunción son obligatorias antes del marcador.

Por otro lado, para que un marcador ordenador cumpla la función discursiva de continuidad en una secuencia, debe estar acompañado de algún otro marcador del conjunto.

Hipótesis 2: Adverbios y marcadores de continuidad

Para que un marcador discursivo de continuidad que a la vez es adverbio («igualmente», «asimismo») se comporte como un ordenador debe aparecer:

- Entre comas si es intraoracional;
- Si ocupa el primer lugar en una oración debe venir seguido de coma (a excepción de «igualmente»);
- El resto de los ordenadores de continuidad de este grupo, solamente aparecen al comienzo de una oración y deben ir seguidos de coma.

Hipótesis 3: Marcador entre signos de puntuación

Para que el marcador cumpla la función discursiva debe aparecer:

- Entre comas si es intraoracional;
- Si ocupa el primer lugar en una oración debe ir seguido de coma.

Hipótesis 4: Marcador que no ocupa el último lugar de una frase o proposición

Para que cumpla la función discursiva, nunca puede ser el último término de una oración o proposición.

Hipótesis 5: Marcador que exige signo de puntuación previo

Para que cumpla la función discursiva, debe aparecer:

- Una coma o punto y coma previo a él si es intraoracional;
- Si ocupa el primer lugar en una oración, la puntuación puede faltar.

Hipótesis 6: Marcador que exige signo de puntuación posterior

Para que cumpla la función discursiva, debe aparecer:

- Una coma luego del término si es intraoracional;
- Si ocupa el primer lugar en una oración, la puntuación puede faltar.

Tomando como base estas hipótesis, se desarrolló una serie de reglas de detección en el programa Xfst, las cuales se detallan en el parágrafo siguiente.

4. IMPLANTACIÓN EN MÁQUINA

Xfst requiere de un archivo fuente, en este caso en *notepad*, donde se declara la información lingüística, que está separada de la máquina algorítmica. En una primera etapa del trabajo computacional, se desarrollaron reglas para el etiquetado de marcadores discursivos con la etiqueta general «MARCD». Posteriormente, se elaboraron diversas reglas, con el propósito de clasificar a los marcadores incluidos en el corpus, de acuerdo con la clasificación de Martín Zorraquino y Portolés.

4.1. Primera etapa del trabajo

En primer lugar, fue necesario definir caracteres, espacios, tabulaciones y todo aquello que pudiera estar al comienzo o al final de cada palabra:

```
define otro1 [ 0 | " " | %’ | %" | %( | %: | "\n" | "\t" | %> | "." ]
define otro2 [ " " | %’ | %" | %) | %: | "\n" | "\t" | %< | "." ]
```

Las barras (|) separan las distintas posibilidades, el signo de porcentaje (%) indica que el signo que aparece pegado a su derecha debe tomarse en su significado original y no confundirse con una notación específica del programa, y las comillas (" ") señala el espacio que hay entre palabras. OTRO1 incluye los caracteres que pueden estar al comienzo de una palabra, incluyendo el “0” que es vacío, cuando una palabra comienza el texto sin siquiera haber un espacio antes. OTRO2 agrupa los caracteres que pueden estar al final de la palabra.

En segundo lugar, se declararon las expresiones que se correspondían con los marcadores discursivos, junto con los signos de puntuación que, de acuerdo con las hipótesis de Prada, los delimitaban. He aquí ejemplos:

```
define MARCD      [ . “ ” En “ ”  primer “ ” lugar %, | %, “ ” en “ ” primer “ ”
lugar %, | %, sin “ ” embargo %, |%, “ ” no “ ” obstante %, |%, “ ” de “ ” todos
“ ” modos %, |%, “ ”es “ ” decir %, | %,  por “ ” lo “ ” tanto %, ]
```

Se puede apreciar, para los casos de marcadores ubicados al inicio de la cláusula, que se tomó como apoyatura el punto de fin de oración de la cláusula anterior para poder localizar los marcadores que estaban al inicio de la siguiente.

Finalmente, en tercer lugar, se dan las instrucciones para el balizado:

```
define FST [ MARCADOR ] @-> [%< M A R C A D O R %> " "] ... [" " %< M
A R C A D O R  %/ %>]  || otro1 _ otro2
```

Estas reglas fueron probadas en un corpus textual de 100.000 entradas constituido por artículos periodísticos de la prensa argentina. A continuación se presentan ejemplos de los marcadores reconocidos:

(…) **<MARCADOR> De todos modos, <MARCADOR/>** afirmó que todavía no puede anticipar cuándo le darán el alta. (…)

(…) **<MARCADOR> Sin embargo, <MARCADOR/>** los funcionarios que lo visitaron ayer afirmaron que lo mantienen alejado del teléfono. (…)

(…) **<MARCADOR> No obstante, <MARCADOR/>** no se espera una reacción oficial del Gobierno israelí (…)

(…) Eso requiere de comprobación **<MARCADOR>, en primer lugar, <MARCADOR/>** y además los nombres corporativos que circulan como eventuales interesados en el Correo (…)

(…)**<MARCADOR> Es decir, <MARCADOR/>** utilizar recursos con los que ya se cuenta, pero sin modificar el flujo futuro

(…) Y**<MARCADOR> por lo tanto, <MARCADOR/>** de cómo un país es o se parece a una Nación. (…)

Los resultados arrojaron un 100% de precisión y un 99% de cobertura.

4.2. Segunda etapa del trabajo

En la segunda etapa del trabajo, opté por crear diferentes reglas de balizamiento, una para cada tipo de marcador. Esto implicó la creación de varios archivos fuente. Se ejemplifica con parte de los archivos elaborados para la detección de marcadores discursivos operativos de concreción (I), conectores contraargumentativos (II) y reformuladores recapitulativos (III):

(I)

```
define MARCOPCON [ %, “ ” por “ ” ejemplo %, | . “ ” Por “ ” ejemplo %, | %,
“ ” en “ ” particular “ ” ];
```

(II)

```
define MARCONCONTRAARG [%. “ ” No “ ” obstante %, | %. “ ” Sin “ ”
embargo %, ];
```

(III)

define MARCREFREC [%, “ ” en “ ” realidad %, | %, “ ” de “ ” hecho “ ” %, | %, “ ” en “ ” suma %,].

Si bien se incrementó el trabajo computacional y este se tornó menos económico, los resultados ganaron en especificidad, al arrojar, en el análisis, el tipo de marcador. He aquí ejemplos de los nuevos balizados:

> A veces, la perturbación moral está determinada por la estructuración de sistemas que, aunque de origen político, afectan la estructura cultural de toda la sociedad y no sólo la conducta de los dirigentes. Es lo que ha ocurrido**< MARCOPCON >, por ejemplo,</ MARCOPCON >** con el manejo abusivo que se ha hecho, en ciertos casos, de la tendencia al reparto de dádivas sin controles ni contrapartidas, en el contexto de un mal entendido asistencialismo social.
>
> (…)
>
> De ningún modo se pretende aquí desconocer las angustias que generan la desocupación, la exclusión y la consiguiente marginación. Lo que estamos tratando de señalar es el efecto pernicioso, en el largo plazo, de aquellos sistemas que deterioran la cultura del esfuerzo personal. Es **<MARCREFREC>, en definitiva,</ MARCREFREC >** lo que ocurre con los planes Trabajar, que a pesar de su engañosa denominación lo que menos alientan es el trabajo.
>
> (…)
>
> **< MARCONCONTRAARG >No obstante, </ MARCONCONTRAARG >**no se espera una reacción oficial del Gobierno israelí pues la noticia se difundió cuando ya había comenzado la festividad judía del "Sabbath", que concluye mañana, sábado, al anochecer. Participación argentina. La Argentina presentó junto con la Unión

Se mantuvieron los porcentajes de la primera etapa y los marcadores detectados fueron clasificados correctamente en el 100% de los casos.

5. CONSIDERACIONES FINALES

El análisis automático de los marcadores discursivos es una de las tareas fundamentales de la lingüística computacional. La correcta detección de estas construcciones puede ser útil

al resumen y la traducción automáticos, o al análisis sintáctico computacional, entre otras cosas. Uno de los problemas, en este ámbito, está dado por los casos de ambigüedad.

A tales efectos, elaboramos un método de etiquetado de marcadores discursivos con la herramienta de estados finitos Xfst, tomando como base la clasificación de Martín Zorraquino y Portolés (1999) y las hipótesis de desambiguación por medio de los signos de puntuación, especialmente de la coma, propuestas por Prada (2001).

El trabajo con Xfst se dividió en dos etapas. En la primera de ellas, etiquetamos los marcadores discursivos con la designación general «MARCD» y, en la segunda, elaboramos varias etiquetas acordes con la clasificación de Martín Zorraquino y Portolés.

Pudimos apreciar que, en ambas etapas, los resultados arrojaron un100% de precisión y un 99% de cobertura. Asimismo, debe señalarse que los marcadores detectados en la segunda etapa fueron clasificados correctamente en el 100% de los casos.

El trabajo a futuro se organiza en torno a los siguientes ejes:

- Continuar con el análisis de los signos de puntuación haciendo hincapié en la manera en que estos ayudan a la desambiguación.
- Utilizar el método de detección de los marcadores discursivos propuesto en tareas de resumen automático.

Referencias

AÏT-MOKHTAR, Salah (1998), *L'analyse présintaxique en une seule étape*. Tesis doctoral. Clermont-Ferrand, Universidad Blaise-Pascal/GRIL.

ALCOBA, Santiago (2000), «Puntuación y melodía de frase», en S. ALCOBA (coord.), *La expresión oral*, Madrid: Ariel Practicum, pp. 147-186.

ALONSO ALEMANY, Laura, CASTELLÓN MASALLES, Irene y PADRÓ CIRERA, Lluís (2003), «Lexicón computacional de marcadores del discurso», *Procesamiento del lenguaje natural* 29: 239-246.

BEESLEY, Kenneth R. y KARTTUNEN, Lauri (2003), *Finite State Morphology*, Stanford: CSLI Publications, Stanford University.

KNOTT, Alistair (1996), *A Data-Driven Methodology for Motivating a Set of Coherence Relations*, tesis doctoral, Edimburgo: Universidad de Edimburgo.

KNOTT, Alistair y DALE, Robert (1995), «Using linguistics phenomena to motivate a set of coherence relations», *Discourse Processes* 18(1): 35-62.

FIGUERAS, Carolina (2001), *Pragmática de la puntuación*, Madrid: Octaedro.

GUILLOT, Daniel E. (2005), «Lingüística computacional: del prototipo a la aplicación», en V. CASTEL, (comp.), *Desarrollo, Implementación y utilización de modelos para el procesamiento automático de textos*, Mendoza: Facultad de Filosofía y Letras, Universidad Nacional de Cuyo.

KOZA, Walter A. (2011), *Los signos de puntuación en el análisis automático de textos. El caso de la coma*, tesis doctoral, Rosario: Universidad Nacional de Rosario.

KOZA, Walter A. «Marcadores discursivos del español. Descripción y propuesta de detección automática», *Revista de Epistemología y Ciencias Humanas* 2: 109-120. Disponible en http://www.revistaepistemologi.com.ar/biblioteca/11.KOZA.pdf [Consultado el 10 de enero de 2012].

MARCU, Daniel (1997), «The rethorical parsing of natural lenguage texts», *ACL-97*: 96-103. Disponible en http://acl.ldc.upenn.edu/P/P97/P97-1013.pdf [Consultado el 10 de enero de 2012].

MARTÍN ZORRAQUINO, M. A. y PORTOLÉS, J. (1999), «Los marcadores del discurso», en I. BOSQUE y V. DEMONTE (eds.), *Gramática Descriptiva de la Lengua Española*, tomo III, Madrid: Espasa Calpe, pp. 4051-4203.

NUNBERG, Geoffrey (1990), *The linguistics of punctuation*, Stanford: CSLI Lecture Notes.

PORTOLÉS, José (1998), *Marcadores del discurso*, Barcelona: Ariel.

PRADA, Juan José (2001), *Marcadores del discurso en español. Análisis y representación*, tesis de maestría, Montevideo: Universidad de la República.

REAL ACADEMIA ESPAÑOLA (2005), *Diccionario panhispánico de dudas*, Madrid: Santillana.

SIMONE, Raffaele (1991), «Riflessioni sulla virgola», en M. ORSOLINI y PONTECORVO, C. (comps.), *La costruzione del texto nei bambini*, Florencia: La Nuova Italia, pp. 221-231.